DOCUMENT RESUME

ED 093 990 TM 003 846

AUTHOR	Fremer, John
TITLE	Developing Tests for Assessment Programs: Issues and Suggested Procedures.
INSTITUTION	Educational Testing Service, Princeton, N.J. Center for Statewide Educational Assessment.
SPONS AGENCY	Ford Foundation, New York, N.Y.
PUB DATE	74
NOTE	36p.
EDRS PRICE	MF-$0.75 HC-$1.85 PLUS POSTAGE
DESCRIPTORS	*Educational Assessment; Evaluation Methods; *Test Construction; *Testing; Test Selection

ABSTRACT

This paper attempts to provide practical guidance to those individuals responsible for selecting or developing instruments for assessment programs. The question of what to measure in an assessment program is addressed at a global and a specific level. Once a developer has identified the areas to be assessed, it is necessary to consider the reporting plans for the program. Whether reporting is by group or individual, the nature of the reporting planned and the types of instruments needed to accomplish it need to be considered. After assessment areas are selected, appropriate instruments need to be selected. Some helpful sources (listings and evaluations of existing tests) are given. In the development of new assessment instruments, six areas are to be considered: initial planning and allocation of responsibility, development of instrument specifications, item development, pretesting, use of item analysis, and final test assembly. Each of these areas is considered at some length. (Author/RC)

CSEA

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS COPYRIGHTED MATERIAL HAS BEEN GRANTED BY
John Fremer
TO ERIC AND ORGANIZATIONS OPERATING UNDER AGREEMENTS WITH THE NATIONAL INSTITUTE OF EDUCATION. FURTHER REPRODUCTION OUTSIDE THE ERIC SYSTEM REQUIRES PERMISSION OF THE COPYRIGHT OWNER."

Developing Tests for Assessment Programs: Issues and Suggested Procedures

John Fremer

CENTER FOR STATEWIDE EDUCATIONAL ASSESSMENT
EDUCATIONAL TESTING SERVICE · PRINCETON, NEW JERSEY

ED 093990

DEVELOPING TESTS FOR ASSESSMENT PROGRAMS: ISSUES AND SUGGESTED PROCEDURES

John Fremer

Educational Testing Service

Published by the Center for Statewide Educational Assessment
which is supported by funds from the Ford Foundation

TABLE OF CONTENTS

Overview 1

What Should Be Measured? 1

What Types of Reports Will Be Needed? 4

Should Newly Developed or Existing Instruments Be Used? 7

How Should New Assessment Instruments Be Developed? 10

- Initial Planning and Allocation of Responsibility 10
- Development of Instrument Specifications 13
- Item Development 16
- Pretesting 20
- Use of Item Analysis 23
- Final Test Assembly 27

Final Comment 29

Developing Tests for Assessment Programs: Issues and Suggested Procedures

Overview

It would be difficult to overestimate the importance of selecting or developing tests, questionnaires, or other measurement instruments that will fulfill the goals of an assessment program. All a measurement instrument can do is permit the systematic collection of information. It is the job of planners and developers of assessment programs to insure that the information obtained is the kind of information that will be helpful in evaluating and making decisions about the status of education in a school district or state. Among the issues that need consideration are the following:

What should be measured?

What types of reports will be needed?

Should newly developed or existing instruments be used?

How should new assessment instruments be developed?

This paper addresses each of these issues in turn, and attempts to identify the major factors that will require attention and to offer possible design and development strategies.

What Should Be Measured?

The question of what to measure in an assessment program is one that has to be addressed both at a global and a specific level. Considering the global level first, one possible answer is that the program should assess the extent to which students, teachers, administrators, and other educational personnel, in short, the entire educational system is achieving the goals for education in a school district or a state. Most states and many school districts already have goal statements that have undergone a cycle of development and refinement. This process can be a very valuable one, particularly if

parents and other members of business and community groups contribute to the task of setting and reviewing overall educational goals and establishing priorities. The participants in the goal setting process are likely to become aware of the extraordinary breadth of goals that schools are being asked to address. These same participants might be able to serve as spokesmen for an assessment program that attempted to measure a wide variety of goals. Even in advance of a systematic review of goals for a school system or state it is possible to make a fairly accurate prediction of the outcome of this review. A recent Cooperative Accountability Project report on State Goals for Elementary and Secondary Education (Zimmerman, 1972), for example, reveals considerable consistency among the goals statements of 35 states. Basic skills goals appear in many forms in the lists developed in the various states, just as they appear, without exception, in the goals for individual school systems.

The assessment program planner and developer has to go beyond the global level and has to deal with the problem of determining the priorities to be assigned to measuring the many goals for education. What emphasis should be given to the basic skills area, to other school subjects, to competencies that have particularly high survival value in our society, and to values and attitudes or other noncognitive attributes of students or of teachers? To what extent should the process of education, e.g., teaching styles, methods of classroom organization, etc., be described and documented? These are difficult questions, moreover, they are not ones that should be answered wholly or even primarily by a technical assessment group. Many educators and members of the larger community have perspectives that will need to be brought to bear on the problems. It is clear, though, that the breadth of educational goals will require a sequential approach to assessment program development. The developer will have to start with some obviously important goal areas, such as reading or communication skills or health or mathematics, and concentrate his initial time and resources on adequate measurement of them. At the same time, long-term plans can be developed for addressing the other significant goal areas.

Assessment program developers have often initiated their programs with testing of reading and mathematical skills at one or more grade levels. Since these skills have high survival value and since a number of measurement instruments and approaches are available, this seems a quite reasonable way to start up a program. More difficulties can be expected if the program developer attempts to assess student or teacher attitudes and values, yet these noncognitive attributes are valued highly by educators, legislators, and private citizens. In order to create an assessment program that adequately reflects the goals for education in a school district or state, some measurement in noncognitive areas is recommended at the very beginning of the program. Awareness of measurement difficulties will encourage postponement of attention to the noncognitive areas. In this connection, it is worth considering the observation of Campbell, Bruno, and Schabacker (1972 p.3): "Although these noncognitive areas are admittedly more difficult to measure, in an assessment program they must not be ignored in the early phase or they most likely will continue to be neglected as the program is enlarged."

When an assessment program developer is making plans for the initial assessment years, attention needs to be given also to the future direction of the program. It will often be difficult to predict the level of funding that will be available, but estimates will have to be made and long-term emphases identified. What goal areas can be added to the program in future years? What kinds of assessment cycles should be introduced? Should some goal areas be assessed yearly and others on an every other year or every third year basis? Which tests can be reused in subsequent years, with or without some revision? Should some provision be made for workshops or special training materiuls for the users of assessment results? Questions such as these go beyond the initial question of what should be measured but they set the stage for the issues addressed in the balance of this paper.

What Types of Reports Will Be Needed?

Once a developer has identified the areas to be assessed in the initial phases of a program, it is necessary to consider the reporting plans for the program. This job should be tackled as early as possible rather than left, as it often has been, until many other decisions about an assessment program have already been made. Decisions regarding the information to be collected and reported will directly affect instrument planning. Is it necessary, for example, to develop reports for individual students? If so, every student must sit for any tests for which such reporting is required. If, on the other hand, reporting will be done for groups of students, sampling procedures such as those outlined by Trismen (1972) and Jaeger (1973) can be employed

Whether reporting is by group or individual, the nature of the reporting planned and the type of instruments needed to accomplish it need to be considered. The state-wide and school district testing programs that are the forerunners of today's developing assessment programs report summary scores and sometimes subscores based on survey, norm-referenced achievement tests. The summary scores can be used to relate state or district results to national norms or to monitor the performance of groups over time. Such summary score reporting has received, however, a great deal of criticism on the grounds that it does not tell us what we need to know in order to take constructive educational action. A good deal of attention has been paid recently to the possibility of reporting assessment results for cognitive areas in terms of specific student competencies, such as the abilities to:

-- address a business letter
-- pass a state driver examination
-- figure correct change, or
-- choose a nutritionally balanced meal

This same logic could also be used to call for reporting on attitudes or behaviors, such as:

-- the number of nonrequired books of various types students read
-- the importance students attach to various rights expressed in Bill of Rights, or
-- the value students at specified grade levels place on certain environmental conditions

The calls for objectives-referenced or content-referenced or criterion-referenced tests have suggested that test developers need to determine precisely what students know or can do. Holders of this position indicate that critical objectives must be identified, and associated measurement procedures developed along with judgmentally or empirically derived standards. These standards would permit a determination of whether or not students had achieved the objectives. Educators can then direct their efforts at those high priority objectives that students have not attained. The argument has typically been framed in a way that calls for measurement procedures that yield only "yes, he has" or "no, he has not" decisions regarding attainment of objectives, e.g., Robert can identify the main idea in reading selections of a specified difficulty level. The approach is easier to defend, however, if the concept of degrees of attainment of objectives is employed and if the probabilistic nature of measurement is kept in mind, e.g., John can type 70 $\pm$ 10 words per minute. Some advocates of objectives-referenced or criterion-referenced measurement have caused educational mischief by seeming to seek the unattainable goal of error-free measurement and thereby creating confusion regarding appropriate standards for measurement instruments, e.g., adopting the untenable position that reliability and validity are concepts which are not applicable to criterion-referenced tests. There have been problems also with the setting of performance levels that will be taken as evidence that a student has attained an objective. Too often, arbitrary levels such as 85% or 95% correct have been

used. Ideally, performance levels would be set with reference to some future situation such as the subsequent educational experiences that are planned for the student. Criterion-referenced testing would then indicate whether the student had achieved the skills and competencies necessary to perform well in the next program or unit of instruction.

The positive effects of the criterion/objectives-referenced testing movement, however, far outweigh the negative ones. One highly significant and positive outcome is that a comprehensive reevaluation of the purpose and uses of tests has been initiated. Developers of testing and assessment programs have had to consider carefully the types of information they can and should obtain from tests and to broaden their thinking about methods of reporting information to the various audiences for assessment results. For a discussion of reporting as it relates to criterion-referenced assessment programs, see "Developing a Criterion-Referenced Assessment Program" (Fremer, 1973).

Some assessment program developers have chosen to make use of the National Assessment pattern of reporting results on an exercise-by-exercise basis. This approach can be employed with any exercise or item and it does seem to stimulate public interest. It is necessary, however, to contend with the problem of overinterpretation. It is natural for readers of such assessment reports to assume that the results from a single question provide insights that can be generalized to whole classes of skills and knowledge. Yet the results from another question tied to the same objective might well be dramatically different and thus lead to different conclusions. Careful pretesting of groups of similar questions can help. Items selected for reporting can be ones with difficulties representative of the total group of items tied to an objective. Even when an item is chosen on this basis, however, the pool of available items may not represent adequately the pool of items that could have been written to measure the objective. It will always be necessary to recognize that measurement and interpretation involve errors and inferences that can lead to unwarranted conclusions. Qualified rather than absolute statements

should be the goal of assessment program developers.

Reports of the proportions of students achieving specified educational objectives perhaps form a middle ground between total score reporting in terms of norms and the reporting of results on individual test items or exercises. (Reports for individual exercises for any given group can, of course, be related to the results for these exercises when administered to some norms group.) Reporting of the proportion of students achieving specified objectives can be the outcome of the administration of homogeneous sets of items or exercises aimed at these objectives. This work or task sample approach has been the typical route to reporting by objectives. It is also possible, though, to use available survey achievement tests to make estimates regarding the proportions of groups of students that have attained specific objectives. The results of survey achievement tests need to be related by experimental procedures to the behaviors or competencies that are of interest. Such an approach involves the use of information on a number of aspects of subject-matter mastery to estimate mastery of particular skills. This idea is developed in a report entitled "Criterion-Referenced Interpretations of Survey Achievement Tests" (Fremer, 1972).

Should Newly Developed or Existing Instruments Be Used?

A developer that has selected assessment areas and decided to use particular types of instruments and reporting approaches will almost certainly have made these decisions with some reference to his knowledge of available instruments and his estimate of the feasibility of developing new instruments. Regardless of the areas chosen, there are likely to be some instruments that would have a claim to appropriateness on the basis of their titles or descriptions appearing in journals or publishers' catalogs. In the area of reading, for example, the Test Collection at Educational Testing Service had collected some 700 tests as of November, 1973 from all parts of the country and the world. Whatever grade level was planned for the testing of reading skills, a stack of

tests of mixed origins and quality could be identified. Knapp (1972, 1973) has provided an indication of the availability of instruments in the non-cognitive areas of school-based attitude and self-concept. Other sources provide listings and evaluations of existing tests. The following are some helpful sources:

Mental Measurements Yearbook Series (Gryphon Press, Highland Park, New Jersey

The volumes in this series include description of tests, critical reviews, publishers' directories, and bibliographical references.

1. Mental Measurements Yearbooks (MMY)
2. Tests in Print
3. Reading Tests and Reviews
4. Personality Tests and Reviews

CSE: Elementary School Test Evaluations and CSE-ECRC Preschool/ Kindergarten Test Evaluation

These volumes include ratings of tests on a number of criteria. They are published by the Center for the Study of Evaluation and the Early Childhood Research Center, UCLA Graduate School of Education, Los Angeles, California.

NCME Measurement News

This newsletter of the National Council on Measurement in Education contains general articles on testing issues as well as announcements of new tests and lists of test reviews.

Test Collection Bulletin (TCB) -- ETS, Princeton, New Jersey

This is a quarterly digest of information on tests and services which generally have become available after the publication of the most recent Mental Measurement Yearbook. It describes both commercially available tests and tests used experimentally. The Bulletin does not evaluate the tests listed, but it does provide references to test reviews.

The abundance of existing tests places a burden on the developer in that attention needs to be paid to their evaluation. In this connection, a committee of reviewers representing the groups who contributed to the goal setting process can be helpful. It is likely to be the case that no existing instrument would be ideally appropriate for any given assessment program; yet, the best available instrument may be judged acceptable, particularly if time, staff, and budgeting constraints permit no other alternative. The use of nationally standardized tests may still be appealing even when the schedule and budget would permit local development efforts, as the fact that standardized tests have had extensive editorial and subject-matter reviews and careful pretesting can be of value in defending a program.

The items in such tests could be matched to educational objectives and reporting carried out for appropriate clusters of test items. It should be recognized that monies not used for development in one assessment area can be allocated to other areas. Use of a standardized reading or mathematics test could therefore free up funds for work in attitudinal or other non-cognitive areas. Ideas for new approaches to testing in either cognitive or non-cognitive areas could be explored and perhaps carried to the point of pre-testing. It would also be possible, for example, to supplement an existing standardized test with newly developed materials covering aspects of content not emphasized in the best available standardized tests. A set of questions on aspects of arithmetic important to twelfth graders as prospective consumers, renters or buyers, and income tax payers could be added to a conventional survey mathematics test. Questions on local or state history or government can supplement the content of a more global Social Studies test.

Whatever the balance of existing or newly developed materials included in an assessment program, it will be desirable to provide some time for pre-testing of new material. It is often necessary to fight for blocks of testing time, and teachers and administrators are understandably reluctant to add to the minimum that was granted during the first year of a program. Failure to seek enough time for the tryout of materials, though, can remove a convenient mechanism for gradual evolution of a program.

How Should New Assessment Instruments Be Developed?

The instruments used in assessment programs and their methods of development are likely to receive a great deal of critical attention from educators, school board members, legislators, private citizens, and the press. It is important, therefore, for program developers to adhere to high measurement standards in the design and implementation of the assessment program. This goal will most likely be achieved if staff can be identified and utilized who have both extensive training in measurement and statistics, and first-hand experience with the development of testing and assessment programs. A school district or state assessment team can include some relative newcomers to the field of assessment, but it must have a solid core of old hands.

Any project is likely to succeed or fail on the basis of the quality of the staff who are running it, yet even a good team is not sufficient. The odds that a good staff will do a good job will be heavily influenced by the extent to which adequate planning takes place. This paper identifies general areas of assessment program development that will require careful thought. Each assessment situation will present its own special problems, but to ignore any of the general issues listed is, in the judgment of this writer, to court trouble. The points to be considered are grouped into the following six areas:

Initial planning and allocation of responsibility
Development of instrument specifications
Item development
Pretesting
Use of item analysis
Final test assembly

Each of these areas is considered in turn.

Initial Planning and Allocation of Responsibility

1. Identification of all components of the instrument development project -- This step involves the participation of all project staff, supplemented by external consultants with skills that round out talents of the project team. An extremely useful

source of information in this connection is the chapter "Planning the Objective Test" by Sherman Tinkelman in Educational Measurement, edited by Robert L. Thorndike.

2. Development of a schedule for the completion of the steps -- This task is most readily accomplished by working backward from identified administration and reporting dates. The length of time needed to accomplish each step is determined using whatever sources are available. The identification of critical sequences can often be facilitated through the use of PERT charts (Wagner, 1973) or other diagnostic or tabular methods of presenting data.

3. Fixing clear lines of responsibility -- The overall Project Director will assign responsibility for aspects of the work to his staff on the basis of their experience and competencies. It will be valuable to not only establish clear lines of final responsibility, but to provide a second or back-up person for every task. The back-up person would review the primary person's work and remain sufficiently involved so that he could step in temporarily in the event of staff changes, illnesses or the like. The use of the Project Director as the only back-up person is to be avoided wherever the size of the staff exceeds perhaps five people. A written statement of responsibilities will be useful for large working groups. Such a statement can help other departments or agencies work efficiently with the project team.

4. Relationship to long-term goals -- Long-term goals usually receive a good deal of attention in the course of making initial program decisions, such as the identification of goal areas for early assessment. It is difficult, however, to continue to keep the long-term goals in mind when making the many specific decisions that design and implementation of a program require. Members of the project team can try to raise questions of long-term impact when they review their own work and that of their colleagues. An advisory group can also be

helpful, particularly if an evaluation of the relationship of present plans to future goals is made part of their charge.

5. Possibilities for multiple uses of assessment program data -- Most assessment programs are developed with more than one use for the data in mind. Emphases do vary, and one program will be focusing primary attention on the provision of global information to administrators, whereas another program will be devoting primary attention to the evaluation of particular programs. It will often be possible to serve an overall major goal quite effectively and still make provision for additional uses of the resultant data. The two examples of global evaluation at the school district or state level and individual program evaluation, for example, are quite compatible. It is true that the program evaluator will need to compare the content of an instrument used in the assessment to the objectives of the particular program, but the assessment program developer can help by providing ready access to the considerations influencing the instrument development process. The local evaluator can be further assisted if the instrument administration pattern produces individual scores that can be aggregated in various ways at the local level.

6. External control of aspects of a program -- The Program Director of an assessment program will want, generally, to maintain the level of control permitted by his position in an administrative system. Consideration should be given, however, to delegating to an external group such as an advisory committee, responsibility for certain components of the assessment. In the development of instrument specifications, for example, a committee of educators might be given a decision-making as opposed to advisory function within limits defined by the Project Director and his staff.

Development of Instrument Specifications

1. Involvement of many groups — The specifications for assessment instruments should probably never be developed solely, or even primarily, by an internal staff group. Even when a school district or state has a large assessment staff with many talents and perspectives, the results of its unaided efforts will be judged unacceptable by the significant groups who were not represented in the specifications development process.

2. Early and continuing external involvement -- The later one waits to involve an external group in the assessment process, the more likely it is that the group will resent the possible implication that they are being called in to "rubber stamp" the plans of internal staff. It is difficult to make changes late in the process of instrument development without bypassing desirable review and quality control steps, so the program developer is likely to resist suggestions for change. An advisory group that is involved early in the development process will have the ability to help formulate those aspects of specifications that are easy to identify and to reach agreement about, as well as the ones that result in disagreement and can only be handled through compromise. An advisory group that has worked through this process will be more likely to defend than to criticize the resultant specifications.

3. Covering all types of specifications -- Discussions of test specifications often center narrowly on subject-matter content for cognitive tests. When considering attitudinal or other non-cognitive areas, it is necessary to expand the concept of "content" objectives to cover the classification of behaviors and occasions. For both cognitive and attitudinal instruments, it is also essential to go beyond content specifications to consider such additional categorizers as the following:

    a. Statistical Specifications -- Appropriate statistical specifications or selection criteria for individual

items and for sets of items need to be developed. Item difficulty will be significant if a norm-referenced instrument is being constructed as this statistic will help guide the development of a test that will differentiate among the levels of skill represented in a particular population. Item difficulty information will also be valuable if an objectives-referenced or criterion-referenced test is being developed. In this latter situation, item difficulty can serve as a check on the reasonableness of particular objectives for various grade levels. It will also be useful to assess the degree of agreement among the difficulty levels of items judged to be equivalent measures of the same objective. If the items fail to yield results congruent with expectations, the items may be testing different attributes than those intended. Item to total test or to subscore correlations will be useful as an index of item homogeneity and as a stepping stone to the evaluation of score reliabilities. Since reliability indices permit an estimate of the likelihood that a similar score would be earned on a parallel set of items, this information is essential to an adequate evaluation of any test. It has been suggested that items for criterion-referenced tests should be selected from among those items that are sensitive to instruction (Roudabush, 1973). Even in this situation, though, scores would have to be stable or reliable in the absence of instruction for the results of testing to be meaningful.

b. Question Type Specifications -- A number of practical constraints have led assessment program developers to

rely primarily on paper and pencil, machine scorable question types. Each assessment program developer needs to consider, though, the possibility that other approaches would be more appropriate to the goal area under consideration. Consider, for example, the measurement of writing ability. Objectively scorable item types have been developed and validated against actual writing ability, yet it is clear that writing ability can only be measured directly through exercises requiring writing. Inclusion of actual writing exercises creates a need for professional scoring of exercises, but the likely increase in the credibility of assessment results may well justify the expense. If the writing exercises are administered only to a sample of students the expense of scoring need not be very great.

c. Stimulus Material Specifications -- In addition to reviewing the possibility that a variety of question formats might be feasible, attention should also be given to the use of other-than-written stimulus material for questions. Tapes, films, and slides might be employed with samples of students or with an entire assessment population. Test administrators can be trained to read certain materials, speak certain sounds, or make use of apparatus of various types. Clearly budgetary factors must be taken into account, but an assessment program must provide you with needed information if it is to be of value. Some types of needed information cannot be obtained by the least expensive testing formats.

d. Cultural Values Specifications -- Cultural values are usually thought of as the province of some special area of testing, such as citizenship, if they are thought of at all.

Yet tests do communicate values to students and it is well to consider this fact when designing the test. What provision is going to be made to represent various groups in the test development process? What guidelines re test content will keep attention focused on an appropriate balance of contributions from many different facets of subject matter fields? What values will be implied by the stimuli and questions?

e. Other Specifications -- The foregoing "special" categorizers do not exhaust the list of item and total test attributes that an assessment program developer needs to be sensitive to. They may be helpful, however, as indications of the breadth of concern that is essential to successful program development. Each program developer will need to work with fellow staff members and with outside people to identify the additional areas for which specifications will be needed.

Item Development

1. Specifications first -- Item development is such a difficult, important and time-consuming part of assessment program development that there is a strong tendency to want to begin item writing without adequate attention to detailed program specifications. It is essential, though, to design content specifications that clearly identify what is to be measured, before item development commences. In some instances this may require the elaboration of detailed educational objectives that are implied by, or subsumed under, existing educational goals for a state. Such work on objectives is an essential part of assessment program development when results are to be reported on an objective-by-objective basis. It is a possible, but not mandatory, procedure when more global reporting is intended.

2. Use of existing models -- Item development for assessment programs is often initiated because of dissatisfaction with existing tests and

the items contained therein. It is inefficient, nevertheless, to ignore existing tests as a source of models, or at least ideas, for new items or exercises. Much can be gained by collecting existing tests and taking a hard look at what is or is not desirable about the constituent items. One can then employ in a new test any format or approach that seems suitable, and identify the undesirable features that the new items will be sure to avoid. It is possible to evaluate the extent to which new items are actually better than the existing ones. The new items can be mixed with the "bad" items from existing tests. All tests should be typed on cards or standard forms so there is no clue to origin. Reviewers can then be asked to rank or assign a quality rating to every item. If the new items are indeed better, they should receive more positive evaluations. (This tactic is not recommended for assessment program developers with fragile egos.)

3. Staff for item development -- Every assessment program developer will need to decide who should write and review the needed test items. Are there staff available in the school district or state Department of Education? Can a local school district obtain help for its program from the state Department of Education? Conversely, can the statewide program draw on school districts or colleges for help? What part should "outside experts" from test publishers, research laboratories, or centers play in the process? These are questions that each assessment program developer must answer in the context of the options available to him or her. Whatever the direction taken, though, staff experienced in instrument development must play a major role in the item development process. It is true that there is room for individuals with all levels of prior experience in an item development group, including some staff who are receiving their first on-the-job training in the area. There have to be experienced hands on board, however, if training is to be successful.

4. Training item writers -- How can one go about training item writers? One effective technique is the item writing workshop. A good workshop provides participants with training in the elements of successful item writing and incorporates a goodly amount of actual writing and reviewing experience. Generally, two or more days will be required so that two or three full cycles of item writing, review, and revision will be possible. The actual writing of items is almost always best accomplished by having item writers work independently. Item review and revision, on the other hand, should involve both individual and group work. The group sessions are opportunities to discuss different aspects of items and to explore alternative approaches. The individual sessions permit the most efficient production of materials.

   It is useful if participants can be prepared in advance for productive learning sessions, through the use of background materials that clearly define the item writing task and permit prior familiarization with both terminology and the fundamentals of technique. If possible, trainees should even write some items to the appropriate specifications and bring them to the first training sessions.

   One component of an item writing session should be the attempt to generate ideas for items or exercises that can be further developed at a later time into usable items. This is a strategy that has been employed in the development of exercises for the National Assessment of Educational Progress. Some review of the exercise development procedures employed by the National Assessment is recommended for anyone considering individual exercise reporting: see Finley and Berdie (1970).

5. Types of reviews needed -- A full treatment of the process of item review needs to touch on many different purposes for reviews. Three such purposes are identified below:

a. It is obvious that assessment items need to be appropriate measures of the objectives of interest. To meet this deceptively simple criterion may require considerable statistical work, but it also calls for reviews by individuals thoroughly familiar with the objectives and with the subject matter domain of interest. Such reviewers can certify the appropriateness and accuracy of items, and with guidance from measurement-trained staff can help evaluate the scorability and reportability of exercises that require judgment, i.e., any non-objectively scorable exercise.

b. Despite all the contributions that subject matter specialists can make to item review, yet another review is needed for consistency of style and clarity of expression. This review is best entrusted to a skilled specialist who has the same role for all assessment instruments and items. This procedure facilitates uniformity of format and style. If possible, this same editor/reviewer should also hold responsibility for controlling the readability level of items and of associated directions and explanatory materials. Only if students can understand the tasks posed to them, is it reasonable to view item and test performance as a reflection of their developed competencies.

c. As a final suggestion in this area, developers should consider a possible review role for parents and other concerned citizens, representatives of a crucial audience for assessment program results. If parents are to contribute effectively they should be brought in early in the review process and should be given appropriate background information about program purposes and procedures. Nontechnical reviews of this nature can serve

a valuable public relations function and can bring helpful information on issues such as the importance ascribed to various types of potential assessment material and the offensiveness and controversiality of exercises. Some input on these issues can also be obtained by giving students a role as reviewers. Students can be given an opportunity to comment on items as part of a post-pretesting session, when pretesting is included as a step in item and instrument development.

Pretesting

1. Use some type of pretesting -- Some form of pretesting is a very desirable, perhaps even essential, component of an effective assessment program. As is noted in later sections of this paper, pretesting provides valuable information to the program development staff but it has other benefits as well. Given the careful public scrutiny that can be brought to bear upon assessment program instruments, the protection against faulty items afforded by pretesting is very welcome. State assessment programs are often legislatively mandated with relatively inflexible time schedules, so the first year of a program may have to proceed at a pace that precludes certain types of pretesting. Even in these circumstances, however, some form of item tryout along the lines of those described in this paper is almost always possible. The problem is one of determining what kind of pretesting is possible within time and budget constraints, constraints which will apply also to the school district assessment program. In the limiting case wherein almost no pretesting is possible, two strategies ought to be considered. The first involves the use of items for the initial assessment battery that have already been pretested on a population similar enough to the assessment group so that judgments of item qualities can be made with some confidence. The second strategy is

simply that of treating the first year of the program as a pretest, even if scores have to be reported, in that information gained therein can be used to revise the instruments for use in subsequent years.

2. Developmental trials -- One form of pretesting that requires very little in the way of time and money is item tryouts conducted by the original item writer with a small number of students. In these circumstances the items can be administered on an individual basis and the students can be interviewed by the item writer. This type of pretesting does not lead, typically, to the development of item statistics. Rather, it permits an opportunity for the clarity of wording of questions and directions to be checked by that individual who is most familiar with the intention of the item. The item writer can observe, to the best of his or her ability to do so, what it is that the students seem to be doing when they answer or solve the problem or question. Do they appear to be carrying out the process originally intended by the item writer, or is there some other method of obtaining the answer that is at variance with the objective for which the item is intended? An item, for example, that is designed to require a student to use insight or to synthesize data from many sources, would be judged suspect if students seem to be answering the items solely from factual recall.

Developmental trials provide an excellent opportunity to discover vocabulary or phrasing of questions that is simply too difficult for the age level for which the items are intended. In order to achieve the maximum benefits from developmental trials, it will be desirable to conduct them with students comprising the lower end of the competency range at the age or grade level under consideration. Another potential benefit from developmental trials is an opportunity to obtain a first fix on the amount of time students will need to respond thoughtfully to the test items.

3. Small group trial -- Perhaps the next level of pretesting in terms of time and money required after developmental trials is the "quick and dirty" administration of items and test materials to small groups of students without carrying out the same level of quality of production of test materials as is intended for the final administration. The use of spirit masters or xerography may well be economical here if the number of cases involved is sufficiently small. The small group trials could be limited to an examination of the effectiveness of directions for items in communicating the nature of the task. At a slightly larger level of involvement, the pretesting could incorporate sample items from each of the various types of items planned for inclusion in the final instruments. The items chosen for pretesting should be representative of other items in the domain, including some at the upper limits of complexity and difficulty.

4. Full scale tryouts -- When it is possible to produce test material at about the same level of quality as the final instrument and to try out these items with groups clearly representative of the actual population, it will be very much to the advantage of the test developer to do so. There will always be an interest, of course, in holding down costs, so consideration should be given to methods of pretesting that are efficient. One opportunity to be explored in this connection is the use of a pretest or experimental section that can be added on to the regular battery in an existing testing or assessment program. As was noted earlier, inclusion of a pretest section in the first assessment battery is likely to be easier than trying to add one in subsequent years. When this opportunity exists, the costs of locating an appropriate sample and of setting up the administration conditions can be eliminated. The additional costs for pretesting may still be substantial as it is necessary to develop the items and to arrange for the production of the materials to be included in the experimental section.

5. Trend line pretesting -- For some purposes pretesting may be most useful if it can be conducted on more than one occasion. When attitude measures are being developed, it is often useful to attempt to trace the development of attitudes over the course of the particular age or grade that will be the subject of study. It is often found that at early ages student attitudes simply lack the stability that would make successful attitude measurement possible. It is better to find this information through pretesting than to incorporate it into the final assessment program only to have to explain away a failure to report results. The use of pre- and post-instruction pretesting can also be explored in the cognitive domain. Part of an assessment program might then be focused on those cognitive areas that are known to be sensitive to the types of instruction now being employed in most of the schools in a state or school district. This technique can be employed either for a reporting-by-objectives assessment program or for a global reporting program. For information on the kind of item analytic procedures that might be used in developing items that are sensitive to instruction, see Roudabush (1973). The issue of appropriate item analyses is treated in more detail in the next section of this paper.

Use of Item Analysis

1. Item difficulty -- As was noted earlier, item difficulty information is valuable for both norm-referenced and criterion-referenced test development. Whenever items can be scored right or wrong, as is the case with most multiple choice items in the cognitive domain, item difficulty can be determined. Similarly one can determine the difficulty of sets of items, such as all the items related to a single objective or all the items relating to a single domain, e.g., a total test score. This kind of information is a necessary prerequisite for assessment program developers who need to build equivalent test forms for use in subsequent years of a program.

2. Item correlations -- One of the most useful statistics for evaluating the adequacy of test items is the item-test correlation. This index can indicate to the developer the extent to which any individual item is measuring about the same thing as other items in a cluster or in the total test. The typical values associated with item to total test correlations will vary as a function of the homogeneity of the content covered by the test as well as with the heterogeneity of the group sitting for the test. The developer will have to become familiar with the range of correlations to expect for any given subject matter domain or attitudinal area. One immediate use of the item to total test correlation is to identify those items that require careful editorial examination for possible ambiguities and technical inadequacies. Whenever an item is included with other items in an item cluster or in a total test because it appears on logical grounds to be a member of the same subject-matter domain or noncognitive attribute, a very low positive item to total test correlation or a negative item to total test correlation is an indication that the item is measuring something other than that intended by the developer. Most frequently the developer will discover that the item is being interpreted in a manner not originally expected or that there is some irrelevant characteristic which is preventing the item from functioning as intended. It will often be possible to revise such an item and use it after a repretesting confirms that the problem has been corrected. It will sometimes be the case, however, that an item will prove unrelated to other items for reasons that are not at all apparent even after an intensive study of the content of the item. The inclusion of the item in a test where it will merely be contributing to some total score is to be discouraged. Results for the item, though, may suggest hypotheses about student competencies that can be followed up in experimental studies.

In certain circumstances the results of an analysis of item correlation may suggest that some subset of items should be treated

differently from the remaining items. This outcome is highly likely when item performance is correlated not only with total test score but also with other items that are thought to be measures of the same objective. This procedure can make it possible to assess the homogeneity of items thought to measure the same objectives. If an item is no more highly related to its own cluster than to all items taken together, there is little evidence for thinking that that objective is indeed measured uniquely by the items that seem on logical grounds to be closely related to it. Further evidence for objectives' interrelatedness can, of course, come from the procedure of correlating item cluster scores with other item cluster scores. If sufficient funds are available, factor analytic procedures can also be employed to refine the clusters of items related to individual objectives.

3. Analysis of options -- Test developers will be greatly aided if they employ item analysis programs that indicate the number and relative test performance level of the students choosing each option to a multiple-choice question. This method of analysis permits the ready determination of those answer choices that are acting to depress item to total test correlations, and can often suggest the nature of the ambiguity or misinterpretation that is interfering with the functioning of items. Such analyses may also suggest other questions that would be more appropriate measures of a given objective, and can shed some light on the nature of student misconceptions or problems of interpretation.

4. Development of scales -- The analytic techniques already mentioned can be combined in order to develop knowledge tests with clusters of items related to somewhat independent objectives. They can also be used in the attitudinal area to sharpen measurement of given attitudes, interests, or values. It is, of course, inadvisable to rely solely on statistical data to refine reportable scales in these

noncognitive areas, but statistical data can suggest hypotheses regarding the organization of a student's beliefs and positions which will permit a sharpening of potential scale definitions. A scale defined in this manner, however, will require careful scrutiny to insure that the final collection of items to be reported in terms of a single score do indeed bear a close relationship to each other that is consistent with the developers' understanding of the nature of the attribute being measured. What the assessment program developer has to avoid is a kind of blind empiricism which could lead to the reporting of scores that have no theoretical organization but which "hang together" in only a statistical sense. It ought always to be possible for the developer to state clearly what a high score on any collection of items should mean and what a low score on that same collection of items should mean.

5. Triangulation -- One invaluable aid to the development of scales in the attitudinal domain and to sharpening one's definition of content areas in the knowledge domain is the collection of independent bits of information regarding the same competency or attribute. This procedure which has been called "triangulation" by some writers can involve using more than one type of item to measure an attribute. It can also make use of non-test indicators such as teacher judgments or counts of observable behaviors as one line in the triangulation procedure. Consider for example the possible assessment area "attitude toward reading." Two different types of items, one requiring direct statements from students and the other requiring responses to objective questions, might be employed. In addition, teachers might be asked to judge how positive their students were toward reading and the school library might be asked to maintain records of the extent to which these same students borrowed and read books. If the information contained from these three sources tended to yield similar conclusions regarding individual students, one could be fairly comfortable that attitude toward reading rather than some other attribute had indeed been measured.

Final Test Assembly

1. Components of test assembly -- Final test assembly encompasses activities such as the review, selection, revising, editing, formatting, and organizing of the items or exercises for the instruments of the operational assessment battery.

2. Clarifying final responsibility -- One individual should have primary responsibility for each instrument in an assessment battery regardless of the number of people contributing to the process and whether or not an outside group has contracted for the task. The involvement of a continuing committee working with the assessment program staff is recommended at the time of final test assembly as it is at this point that all earlier work is synthesized. The final test assembler, though, needs to have the authority to make the many decisions which will come up as the test nears completion.

3. Final item review -- What precisely are the tasks facing the responsible individual, his cooperating staff members, and the committee? One significant task is final item review. All information available about the items in the pool should be collected in a convenient form and each item reviewed in the light of this information. One useful strategy in this respect is the preparation of spiral notebooks with items on one page and the following on facing pages:

   - objective or area of specifications covered
   - pretest information (if any)
   - previous reviewers' comments
   - correct response or scoring guide

   Whenever items are to be reviewed, it will be useful to keep the correct response separate from the text of the item so that the reviewer can choose or formulate an answer and then check it against the official key or scoring guide. If the individual with primary responsibility for a test concludes that an item or exercise is

ambiguous or that it lacks a single correct key, that item should not be used in a test, irrespective of the quality of its pretest statistics. Similarly, an item in the attitudinal area that appears to be subject to irrelevant interpretations should not be used as part of a scale, again regardless of its pretest statistics.

4. Meeting assembly specifications -- The process of screening out items because they are judged to be inadequate by reviewers can have the effect of reducing the pool of items in some areas so that it appears impossible to meet the original specifications for a test. At this point it is necessary to consider whether some previously rejected items can be revised, whether additional materials can be created or whether the intended scope of the test will have to be reduced. If it proves necessary to narrow the focus of a test, it will be important to describe just what is and is not being measured by the instrument that is used operationally. In rare cases the screening and culling process may produce a significantly larger body of items than is needed to meet specifications. In such instances, one can sample from the pool in such a way as to leave a set of items that is approximately equivalent to the items used, thus creating the possibility of a parallel form for subsequent use. When a test is designed to show the large variations in competence that are likely to be present in populations, statistical considerations will often help the developer determine which items to use. When both statistical and content dimensions need to be satisfied, few developers will find that there is a large surplus of items in many areas.

5. Coordination with test production staff -- When organizing the final set of items into total tests or sets of related exercises, it will be useful to consult with the staff members who will be responsible for producing many copies of the final test. Decisions regarding the layout of items on pages and the order and sequence of items may have considerable implications for the total cost of producing the final package. No assessment program developer will be comfortable

with page layouts that introduce complexities to questions beyond those necessitated by the nature of the task. The use of type too small to permit easy reading, or excessive packing of questions into pages may undercut the most careful effort to produce quality instruments. Even when consultation with production staff is possible prior to final page layouts, the individual with primary responsibility for an instrument should review the printing masters prior to the test production runs. It is at this point that one is likely to discover such horrendous outcomes as the fact that stimulus and response materials have been inadvertently separated, the options for multiple-choice questions are improperly sequenced, or that no space was provided for students to respond to free response questions.

6. Documentation -- Although the task of test assembly is often so complex and demanding that it is difficult to set aside the time to keep accurate records of decisions made, the absence of such records can often create substantial problems for the program developer. In general, every effort should be made to pick up potential errors as early as possible in the development process, so that last minute changes that will not receive a significant number of later reviews can be avoided. There will always be a need, however, for changes to correct errors that are discovered at the eleventh hour. Careful documentation of the reason for the change, the nature of the change, and the steps that were taken to inform all significant people will reduce the probability of catastrophic errors. Imagine the consequences of rearranging the questions for a test at the last minute, so that the numbers of different items were changed, without notifying the individual who has already prepared the official scoring key for the test.

Final Comment

This paper has attempted to provide practical guidance to those individuals responsible for selecting or developing instruments for assessment programs. The suggestions that have been offered are all

based on first hand experience with the task of developing such instruments, yet it is clear that in any individual situation other possible courses of action could have been suggested; and, if followed, might have yielded quite satisfactory results. There is no one correct way to develop an assessment program, but the enterprise has so many facets that specific suggestions regarding ways that a number of component tasks could be handled may be of value.

References

Buros, Oscar K., Editor. The Seventh Mental Measurements Yearbook. Highland Park, New Jersey: The Gryphon Press, 1972. (The Yearbooks have been published periodically since 1932 and are a comprehensive source of factual and evaluative information about tests.)

Buros, Oscar K., Editor. Tests in Print: A Comprehensive Bibliography of Tests for Use in Education, Psychology, and Industry. Highland Park, New Jersey: The Gryphon Press, 1961.

Buros, Oscar K., Editor. Reading Tests and Reviews. Highland Park, New Jersey: The Gryphon Press, 1968.

Buros, Oscar K., Editor. Personality Tests and Reviews. Highland Park, New Jersey: The Gryphon Press, 1970.

Campbell, Paul B.; Bruno, Nancy L.; and Schbacker, William H. "Statewide assessment: methods and concerns." Princeton, New Jersey: Educational Testing Service, 1972.

Finley, Carmen J. and Berdie, Frances S. "The National Assessment approach to exercise development." Ann Arbor, Michigan: National Assessment of Educational Progress, 1970.

Fremer, John. "Criterion-referenced interpretations of survey achievement tests." Test Development Memorandum 72-1, Princeton, New Jersey: Educational Testing Service, 1972.

Fremer, John. "Developing a criterion-referenced assessment program." Paper presented at annual meeting of the National Council on Measurement in Education, New Orleans, February, 1973.

Hoepfner, Ralph, Editor. CSE Elementary School Test Evaluations. Los Angeles: Center for the Study of Evaluation, UCLA Graduate School of Education, 1970.

Hoepfner, Ralph; Stern, Carolyn; and Nummedal, Susan G., Editors. CSE-ECRC Preschool/Kindergarten Test Evaluations. Los Angeles, California: UCLA Graduate School of Education, School Evaluation Project, Center for the Study of Evaluation, and Early Childhood Research Center, 1971.

Jaeger, Richard M. "A primer on sampling for statewide assessment." Princeton, New Jersey: Educational Testing Service, 1973.

Knapp, Joan. "An omnibus of measures related to school-based attitudes." Princeton, New Jersey: Educational Testing Service, 1972.

Knapp, Joan. "A selection of self-concept measures." Princeton, New Jersey: Educational Testing Service, 1973.

Roudabush, Glenn E. "Item selection for criterion-referenced tests." Paper presented at annual meeting of the American Educational Research Association, New Orleans, February, 1973.

Tinkelman, Sherman N. "Planning the objective test." In R. L. Thorndike (Ed.) Educational Measurement, Washington, D.C.: American Council on Education, 1971, pp. 46-80.

Trismen, Donald A. "Sampling techniques for assessment." Princeton, New Jersey: Educational Testing Service, 1972.

Wagner, Andrew R. "What you always felt you should know about PERT, but were afraid to find out." Research Memorandum 73-15, Princeton, New Jersey: Educational Testing Service, 1973.

Zimmerman, Alan. "Education in focus: A collection of state goals for public elementary and secondary education." Denver, Colorado: Cooperative Accountability Project, 1972.